Measuring Metacognition: The Effects of Framing and Scale Type on Metacognitive Accuracy

Rachel Breen

A report submitted as a partial requirement for the degree of Bachelor of Psychology with Honours at the University of Tasmania, 2017

Statement of Sources

I declare that this report is my own original work and that contributions of others have been duly acknowledged.

Signed_____ Date: 19/10/17

Acknowledgments

I would like to acknowledge the contributions and supervision provided by Dr Matthew Palmer, and thank him for his support and advice. I would also like to thank Jim Sauer for his contribution towards the interpretation of results. Thank you to Laura Brumby for your help programming the study, and for all the advice along the journey. Thank you also to Simon Bury for your assistance with data collection.

I would also like to thank my friends and family for their endless encouragement and support. I would like to extend this thanks to my fellow honours student. Your company and support has been invaluable. Finally, I would also like to extend my heartfelt thanks to all who participated in this study. Your contributions are greatly appreciated.

Table of Contents

List of Tables

List of Figures

Measuring Metacognition: The Effects of Framing and Scale Type on Metacognitive Accuracy

Rachel Breen

Words: 9,712

Abstract

Accurate metacognitive judgements are necessary to predict the likelihood of recalling information and engage in effective learning. The study explored how manipulating JOL question frame and the scale of assessment affected metacognitive accuracy. Ninety-Two participants (59 female) aged 18-69 (M = 27.58, SD = 12.56) completed a cued-recall task of sixty-six English noun word-pairs and made JOLs. The method of JOL assessment (0-100% scale or binary task) and the frame of the JOL question (likelihood of remembering or forgetting) were manipulated to form four between-subjects conditions. The study-JOL-recall procedure was repeated to form two test cycles. Framing had a minimal impact on JOLs when measured on 0-100% scales. For binary conditions, the effects of framing differed depending on test cycle. It was concluded framing effects may result from the selection of different anchors, and that binary tasks may be more sensitive to framing. This study has implications for how health professional, educators, and researchers may consider assessing the beliefs about memory of those they teach. The study highlights the way assessments are made can influence the accuracy of metacognitive beliefs. Hence further research into and development of methods to accurately assess metacognition in both laboratory and real-world settings is necessary.

The accuracy with which people determine whether or not they will recall information in the future can affect learning efficacy (Metcalfe & Finn, 2008). For example, overconfidence in predicted memory performance can result in ceasing study before fully understanding materials (Finn, 2008). Predicting memory performance requires metacognition (Kelly & Donaldson, 2016). Metacognition is the ability to think about one's thinking, encompassing processes such as monitoring comprehension, reflecting on learning, and assessing the efficacy of other cognitive abilities (Kelly & Donaldson, 2016). Judgements of learning (JOLs) are a type of metacognitive task which involves predicting the likelihood of future remembering (Nelson & Dunlosky, 1991;Undorf, Böhm, & Cüpper, 2015).

JOLs may influence study decisions (Metcalfe & Finn, 2008), meaning the accuracy of these judgements can affect learning efficacy. For example, student's JOLs can affect their judgment of which areas need more study and which have been learnt sufficiently (Metcalfe & Finn, 2008). Hence student's may incorrectly believe further study is not required if JOLs are overconfident. The potential negative implications of inaccurate judgements apply in many settings, including judging memory for health information or predicting performance during important educational tests. Improving the accuracy of metacognitive judgements is therefore important to ensure accurate monitoring of information retention and facilitate learning.

JOL accuracy can be assessed by analysing resolution, calibration and over/underconfidence. Resolution (relative accuracy) is the ability to distinguish between items which will and will not be recalled (Zawadzka & Higham, 2015). Calibration (absolute accuracy) is how well mean JOLs match overall memory performance (Zawadzka & Higham, 2015). By assessing calibration, it is possible to

determine if judgements were overconfident (JOL > performance), underconfident (JOL < performance), or accurate (JOL = performance) (Hanczakowski, Zawadzka, Pasek, & Higham, 2013). Over/underconfidence is a measure which indicates how overconfident or underconfident judgements were (Jonsson & Allwood, 2003).

JOLs can be made immediately following the presentation of each item (immediate JOLs) or delayed until after multiple items have been presented (delayed JOLs) (van Loon, de Bruin, van Gog, & van Merrienboer, 2013). Making JOLs after a delay has been found to improve metacognitive accuracy compared to immediate JOLs (van Loon et al., 2013). Differences between immediate and delayed JOL accuracy have been attributed to the type of memory each judgement relies upon. Delayed JOLs likely have greater reliance on long-term memory (Kvavilashvili & Ford, 2014; Schneider, Visé, Lockl, & Nelson, 2000). Hence these decisions can be more predictive of memory performance when recalling information is not required immediately after information presentation (Schneider et al., 2000). This means a simple way to improve JOL accuracy could be to delay judgements.

Although metacognitive accuracy improves following delayed JOLs, this practice may not be applicable in some real-world settings. For example, it would be impractical and time consuming for busy health professionals to assess client's likelihood of remembering information hours after the information was presented. A better method could be to ensure health information has been learnt and understood immediately after it is presented. In doing so, the health professional can either confirm necessary information has been learnt, or to continue to educate the client until understanding has been achieved. Hence this study aims to investigate ways to improve JOL accuracy for judgements made immediately after information presentation.

Manipulating Framing

The impacts of changing question wording on participant's responses has been observed within decision-making literature. For example, Prospect Theory (Kahneman & Tversky, 1979) suggests the way information is framed can impact decision-making. People may view potential resource gains and losses as psychologically different, with potential loses attracting greater psychological weight than potential gains (Kahneman & Tversky, 1979). Decisions framed as potential loses result in the selection of higher risk options. For example, participants may be given $40 and asked to choose between either giving back $10, or making a 50%-50% bet to losing $20 or losing everything. In this condition, participants choose the risky (betting) option. In comparison, framing choices as potential gains results in the selection of less risky options. For example, participants may be given $20 and asked to choose between either being given an extra $10, or making a 50%-50% bet to winning an additional $20 or losing everything. In this case, participants select the first option as it is less risky. Framing decisions as a potential gain therefore increases risk aversion to minimise resource loss. This effect is found despite all options having rewards of the same magnitude (Kahneman & Tversky, 1979). Hence framing may influence decision-making.

Like decision-making, metacognitive accuracy can be affected by question frame (Finn, 2008; Serra & England, 2012). JOLs are typically assessed using a remember frame; that is, participants are asked the likelihood of remembering information in the future (Finn, 2008). Yet changing this assessment from a remember to a forget frame (changing the question from "what is the likelihood you will remember?" to "what is the likelihood you will forget?") may improve prediction accuracy (Finn, 2008). When comparing JOLs assessed with remember or

forget frames, Finn (2008) found improved calibration and less overconfidence when forget frames were used. Finn also found forget frames resulted in more easy and medium difficulty items being selected for restudy, despite equivalent performance between frames on the memory test. The region of proximal learning model (Metcalfe & Kornell, 2005) suggests learning is best when easy and medium difficulty information is mastered before difficult information (Xu & Metcalfe, 2016). Finn therefore suggest selecting these items for restudy indicates more adaptive learning choices. This finding is therefore important, as forget frames may both improve judgements about future memory performance and increase adaptive study behaviours.

Using forget frames to assess immediate JOLs may improve metacognitive accuracy and have positive behavioural consequences (Finn, 2008). Finn (2008) suggested this occurs because framing can promote reliance on different cues to make JOLs. That is, framing changes the metacognitive context and affects decisions about which memory cues to utilise (Serra & England, 2012). Forget frames were suggested to improve JOL accuracy and decreasing overconfidence by heightening sensitivity to cues about memory fallibility and memory decay over time (Finn, 2008). Finn's idea therefore suggests forget frames allow people to monitor learning and information loss more effectively, rather than simply biasing responses.

More recent research indicates the effects of framing may not be as simple as Finn (2008) suggests. Serra and England (2012) also investigated the effects of framing on immediate JOL accuracy, although with some procedural changes. In Finn's study, JOLs were only assessed once. This meant participants studied and underwent JOL assessments for each piece of information, then completed a memory test. Serra and England used a multi-cycle design. A multi-cycle design repeats the

procedure of a single cycle design so that study, JOL assessments, and memory tests are completed multiple times. Use of multi-cycle design revealed remember frames were well calibrated in the first cycle, but became underconfident on the second. Forget frames were overconfident on the first cycle, yet became well calibrated on the second. This conflicts with Finn's results.

Differences in calibration depending on cycle may suggest framing does not alter sensitivity to cues about memory fallibility or memory decay. Instead, Serra and England (2012) theorised framing may promote reliance on different anchors. The anchoring hypothesis (Scheck & Nelson, 2005) proposes an anchor is a value formed during initial JOL assessments. This anchor represents a point on a continuum above which information is considered recallable, and below which it is considered unrecallable (Hanczakowski et al., 2013). The value of this anchor depends on factors such as beliefs about memory when making the initial JOL, and the experimental or task requirements (Scheck & Nelson, 2005). The anchor becomes the base from which later JOLs are made, with people shifting judgement up or down from the anchor depending on their beliefs about future recall (Hanczakowski et al., 2013). Anchoring effects mean JOLs often change little upon retesting (Scheck & Nelson, 2005). This effect is compounded by use of a stability bias; the belief memory is stable over time and will not improve with learning or suffer from forgetting (Kornell & Bjork, 2009).

To test the potential relationship between anchoring and framing, Serra and England (2012) compared differences between JOLs and a second order judgment (SOJ) indicating confidence in the accuracy of each JOL. JOLs move away from anchors as more information about potential memory performance is acquired (Dunlosky, Serra, Matvey, & Rawson, 2005). Likewise, SOJs increase when more

information is available (indicating greater confidence), but remain close to anchors when less information is available (Dunlosky et al., 2005). Lower SOJs can therefore indicate when JOLs are closer to anchors (Serra & England, 2012). Based on difference between JOL and SOJ by frame, it was concluded forget frames result in anchors around the midpoint (50%). Remember frames were suggested to be anchored lower, between 20% and 40%. Thus, differences observed between frames may result from anchoring rather than sensitivity to cues regarding memory fallibility (Serra & England, 2012).

In contrast to both Finn (2008), and Serra and England (2012), there is research suggesting framing does not affect the calibration of metacognitive decisions. Rhodes and Castel (2008) compared metacognitive accuracy for words written in either a small or large font size. JOLs were defined as assessments which used a remember frame. Forget frame decisions were defined as judgements of forgetting (JOFs). JOFs converted to the same scale as JOLs by reverse scoring JOFs. Judgements of remembering and forgetting were found to result in equivalent predicted memory performance. Hence framing was not found to influence metacognitive accuracy (Rhodes & Castel, 2008). It is therefore possible framing does not affect JOL accuracy.

A final finding regarding the effect of framing on metacognition relates to the ability to distinguish between information that will or will not be recalled in the future (resolution). Both Finn (2008) and Rhodes and Castel (2008) found minimal differences in resolution between frames. However, Serra and England (2012) extend on this finding by demonstrating forget frame JOLs may have poorer resolution than remember frame JOLs in addition test cycles. This difference was suggested to occur as forget frames reduce reliance on the memory for past test (MPT) heuristic. The

MPT heuristic suggests people consider performance during initial testing to infer memory performance for the same information during secondary testing (Serra & Ariel, 2014). Reliance on the MPT heuristic has been associated with improvements in metacognitive accuracy, and has been used to explain why resolution improves across test cycles (Finn & Metcalfe, 2008; Serra & Ariel, 2014). As observed by Serra and England, reduced use of the MPT heuristic in cycle two of forget frames would therefore result in poorer resolution in forget than remember frames. As other framing studies have not considered multiple cycles (Finn, 2008; Rhodes and Castel, 2008), further investigation of Serra and England's idea decreased use of the MPT heuristic within forget frames promotes differences in resolution.

Limitations of Framing Research

Framing research is currently limited to a small number of studies. These studies conflict greatly, meaning a pattern of effects has not been established. There is also variation in hypotheses as to why framing effects may or may not occur. While Finn (2008) suggests framing may influence the cues used to predict recall, Serra and England's (2012) research indicates framing may instead alter anchoring. More research is necessary before either theory can be supported.

One way to investigate why framing occurs could be to examine the effects of framing on different scales. There is some research suggesting one method of assessing JOLs is affected by factors altering confidence in recall, while another assessment method is minimally affected by confidence (Hanczakowski et al., 2013). If framing effects were present on both scales, this may indicate framing results from the utilisation of different memory cues. Hence framing may improve the ability to accurately predict recall. Framing effects on only the scale affected by confidence

would suggest framing might result from anchoring effects. As such, this would suggest framing does not improve the accuracy of metacognitive judgements.

The Effects of Scale Type

JOLs are commonly assessed using 0-100% scales. 0-100% scales require participants to rate JOLs on a scale from 0 (no likelihood of remembering) to 100 (absolute likelihood of remembering) (Hanczakowski et al., 2013). Alternatively, JOLs can be assessed using binary tasks. Binary tasks simply requires a 'yes' or 'no' response to a question assessing whether information is likely to be remembered in the future (Hanczakowski et al., 2013).

It has been assumed both scales measure an individual's subjective beliefs about the probability of information recall (Hanczakowski et al., 2013). However, recent research suggests this is not the case. As observed by Hanczakowski, Zawadzka, Pasek, and Higham, (2013) when investigating the underconfidence with practice (UWP) effect, scale type may influence the calibration of immediate JOLs. The UWP effect occurs when JOLs are overconfident on initial testing, yet fail to improve to the same extend memory does during later testing (Zawadzka & Higham, 2015). Interestingly, Hanczakowski et al. found the UWP effect was present when 0-100% scales were used, but absent for binary tasks. This suggests scale type influences JOL accuracy, and may indicate one scale does not assess beliefs about recall probability. More recent research by Zawadzka and Higham (2015) has also found scale type influences calibration, supporting this notion.

Theoretical Accounts of Scale Effects

The cause of scale effects remains unclear. Research by Zawadzka and Higham (2015) suggests variations in JOL accuracy do not simply result from differing scale format. Factors including differences in the number of response options (binary tasks have two response options, while there are multiple options for 0-100% scales) and use of numerical labels in 0-100% scales have not been found to influence JOL accuracy (Zawadzka & Higham, 2015). These factors are therefore suggested not to produce differences in JOL accuracy between scales.

A plausible explanation of scale differences was proposed by Hanczakowski et al. (2013). Scale differences may occur because the scales changed the interpretation of the JOL question. It was suggested either a confidence or probability interpretation could be made. 0-100% scales were suggested to promote confidence interpretations, whereby people make a yes-no decision about the probability of recall and use the scale to make an ordinal ranking of confidence. An item assigned the value of 80% therefore does not indicate an 80% likelihood of recall. Rather, this is a ranking indicating greater confidence in recall when compared to lower ranked items. 0-100% scales may therefore be susceptible to cues which affect confidence in memory such as the UWP effect. Data obtained from 0-100% scales may therefore indicate confidence rather than the probability of recall.

In comparison, binary tasks were suggested to promote probability interpretations. Under these interpretations, answers were suggested to directly index the individual's subjective beliefs about the probability of an outcome. Binary tasks may therefore be a more accurate measure, as the interpretation of JOLs often relies on JOLs representing probability rather than confidence information (Hanczakowski

et al., 2013). This idea may explain why binary tasks were not affected by the UWP effect within Hanczakowski et al.'s (2013) research.

It should be noted that unlike framing effects, scale type may not alter psychological processes. In particular, processes such as the use of certain cognitive biases, or the ability to discriminate between what is and is not known may be unaffected by scale type (Hanczakowski et al., 2013). Instead, it is probable scales simply alter the ability to express beliefs about memory. Hanczakowski et al. (2013) explored this idea by asking participants to make JOLs on a 0-100% scale, immediately followed by making JOLs on a binary task. Participants displayed underconfident JOLs when using the 0-100% scale, but did not show underconfidence when completing the binary task for the same information only moments later. It is unlikely participant's beliefs about the likelihood of recall would change in the few moments between assessments. It is therefore suggested scales alter how decisions are expressed rather than information monitoring.

Limitations of Scale Effect Research

Research on differences between 0-100% scales and binary tasks is currently limited to a small number of studies. Due to this, more research is needed before it is possible to determine whether 0-100% scales, binary tasks, or other methods of assessment may provide the most accurate information. More research is also needed to assess the validity of Hanczakowski et al.'s (2013) theory on differences between probability and confidence interpretations. For this reason, the present study aims to add to this body of research by comparing the accuracy of JOLs made using 0-100% scales and binary tasks.

Aims

One aim of the current study was to investigate how framing may impact metacognitive accuracy. Few studies have investigated framing. In addition, the results of these studies conflict greatly. More research is therefore necessary to establish a pattern of effects. Therefore, the first aim of the present study was to help clarify conflicting research on how manipulating frame may impact metacognitive accuracy.

Investigating the effect of manipulating framing on multiple scales may help achieve this aim. As noted by Hanczakowski et al. (2013), differing interpretations of JOL questions may mean it is inappropriate to suggest a manipulation affects JOL accuracy if impacts are only seen on one scale. Conclusions about the effect of manipulating frame may therefore be inaccurate if the effect on both scales is not considered. As previous studies have considered the impact of framing on 0-100% scales only (Finn, 2008; Rhodes & Castel, 2008; Serra & England, 2012), a second aim of the study was to investigate whether there would be converging results of framing on both scales.

A final aim was to explore possible causes of framing effects. Investigating whether framing differs depending on scales may provide additional information about why framing effects occur. Using Hanczakowski et al. (2013) suggestion 0-100% scales are influenced by factors affecting confidence, framing may be found on 0-100% scales but have minimal influence on binary measures. This may suggest framing affects confidence or changes the anchors used rather than cues about memory fallibility. This would align with Serra and England's (2012) anchoring theory of framing. Framing effects on both scales would support Finn's (2008) suggestion framing affects the cues used in memory decisions (cf. simply altering

confidence in memory). An absence of framing effects would support Rhodes and Castel's (2008) finding framing has minimal effect on metacognitive accuracy.

Hypotheses

The UWP effect was predicted to occur on 0-100% scales but not on binary measures. This would support the research of Hanczakowski et al. (2013), providing additional evidence the interpretation of the JOL question may differ depending on the scale used. The effect of framing may replicate previous research. The finding forget frames promote greater JOL accuracy on both scales would support the research of Finn (2008), suggesting forget frames promote the use of more predictive memory cues. Alternatively, framing may influence anchoring, as predicted by Serra and England (2012). Should this occur, it is expected manipulating frame would have larger impacts on 0-100% scales, as these scales promote confidence interpretations of the JOL question. No effect of framing on either scale would support research by Rhodes and Castel (2008).

Method

Participants

Ninety-five participants completed the study. Two were removed due to missing data. A third participant was removed after reporting incorrect use of the JOL measure, leaving a final sample of 92 participants (59 female) aged 18 to 69 (M = 27.55, SD = 12.57). This conforms with minimum of 20-30 people per cell suggested necessary to reduce the risk of false positive results (Simmons, Nelson, & Simonsohn, 2011). Using G*Power software (Version 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007), a total sample size of 76 was suggested necessary to identify a

moderate effect. Hence the achieved sample size is also consistent with tests of necessary sample size. Twenty-eight participants were undergraduate psychology students from the University of Tasmania. The remaining participants were drawn from the broader Launceston population. Participants were recruited using posters displayed at the Newnham campus of the University of Tasmania and online. Participants either received course credit or were financially remunerated $20 for their time.

Design

A 2 ×2 ×2 mixed design with four between-subjects conditions was used. Scale (0-100% scale or binary task) and Frame (forget or remember frame) were between subjects variables, while Test Cycle (cycle one or cycle two) was a within subjects variable. These four conditions have been referred to as scale-remember (*n* = 21), scale-forget (*n* = 22), binary-remember (*n* = 27), and binary-forget (*n* = 22).

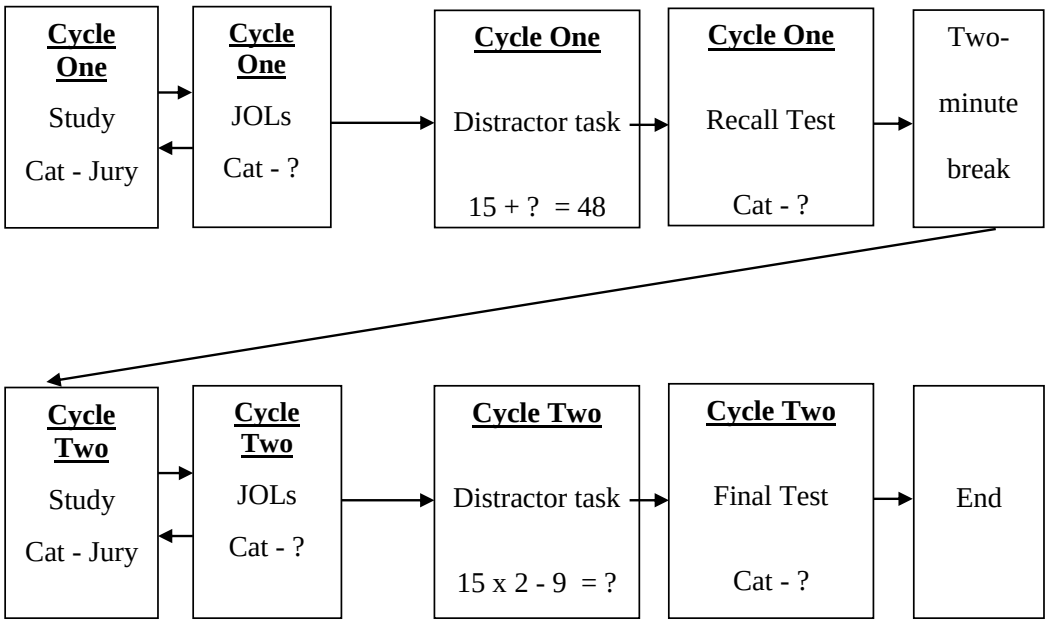


Figure 1: Experimental procedure (multicycle design).

Materials and Procedure

Participants were randomly allocated to conditions using a random number generator. LimeSurvey (Version 2.06; Schmitz, 2015) was used to implement the experiment. Tasks were undertaken individually on a computer. Participants were provided with a verbal overview of the task. They were informed they would be undergoing a memory task which required the learning a list of pairs of words, and later completing a memory test. Participants were verbally informed they would be asked to make judgements about their expected performance during the test for each word-pair, and instructed on how to make these judgements depending on the condition they were in. All instructions were also provided in writing. Information sheets and consent forms were then presented and completed using LimeSurvey (Appendix B). After providing consent, participants began the study. Demographic information including age, sex, current enrolment in first year psychology classes, and whether English was their first language was collected (Appendix C).

The experiment was designed to closely follow the method of both previous framing and scale studies. As can be seen within Figure 1, participants completed a cued recall task over two study-test cycles. Cued recall tasks have previously been used in both framing (Finn, 2008; Serra & England, 2012) and scale effect studies (Hanczakowski et al., 2013; Zawadzka & Higham, 2015). Sixty-six English non-associated word-pairs were presented (Appendix D). These pairs have previously been used within framing research (Serra, Dunlosky, & Hertzog, 2008; Serra & England, 2012). Two randomised lists of word-pairs were constructed. The first list was presented in cycle one. The second list was presented in cycle two. The presentation order of word-pairs was the same for the study and recall phases of each cycle. The presentation and study order of word-pairs was the same for all

conditions. Following the method of Hanczakowski et al. (2013), and Zawadzka & Higham (2015), each pair was presented for 3.5 seconds. The first three and last three pairs were included as buffers to control primacy and recency effects (Rhodes & Castel, 2008). These 6 pairs were not included in the analysis.

JOLs were assessed immediately following the presentation of each pair. Hence a word-pair was presented, the JOL for that pair made, followed by the presentation of the next word-pair. During JOL assessments, the first (cue) word was presented while the second (target) word was replaced with a question mark. JOL instructions were specific to the condition. Instructions were framed in terms of either remembering or forgetting. Differences also existed between scale conditions. Those in the 0-100% scale conditions were asked "*How likely are you are to (remember/forget) the second word in the pair in a few minutes from now?*", where 0% indicated not likely to remember/forget and 100% indicated likely to remember/forget. To respond, participants moved a marker on a bar to indicate a number between 0 and 100. The marker appeared in the middle of the bar (50) for each JOL. This method was selected to remove errors such as entering values outside of the 0-100 range, which could occur if people were able to type responses. It was also considered a more sensitive measure, allowing finer grain distinctions than simply selecting between response options at each 10%. The binary conditions were presented with the statement: "*Are you likely to (remember/forget) the second word in the pair in a few minutes from now*?" Participants responded either "yes" or "no" by clicking the relevant response button. Again, this method was considered superior to typed responses as it prevents typing mistakes. JOL response time was not limited. Participants pressed "next" to move to the next word-pair.

A mathematics filler task was completed for two minutes following the presentation of all word-pairs. Participants were given 18 questions of increasing difficulty (e.g. 3 × ? + 11 = 41) and asked to answer as many as possible. No participant completed all questions in the allocated time. This task was included to prevent item rehearsal before the test period, and ensure working memory was not being assessed.

Following the filler task, participants underwent a recall test. Participants were required to type the target word when the relevant cue word was presented. Participants pressed "next" to move onto the next cue word. Participants could type "x" if they could not recall the correct response. This was scored as incorrect. The study-JOL-test procedure was then repeated to form a second test cycle.

Results

IBM SPSS Statistics version 25 was used to test assumptions and analyse data. The effects of manipulating framing and scales on cue-recall performance and JOLs were examined. Following this, a series of mixed ANOVAs and tests of simple main effects were conducted to examine the effects of manipulations on metacognitive accuracy. As within previous framing studies, forget JOLs were converted to the same scale as remember JOLs by subtracting the JOL value from 100. Partial eta squared has been reported as an indication of effect size for interactions. For pairwise comparisons, Cohen's *d* has been calculated. Values of 0.2, 0.5, and 0.8 represent a small, moderate, and a large effect respectively (Cohen, 1988). Exploratory Software for Confidence Intervals (ESCI) was used to calculate 95% confidence intervals for Cohen's *d* (Cumming, 2012). Descriptive statistics are displayed separately for each cycle in Table 1 and Table 2.

Tests of Assumptions

Z-scores were calculated to identify outliers. Scores above 3.29 were considered potential univariate outliers (Tabachnick & Fidell, 2007). Analyses were rerun, and results with and without outliers compared. As removal of outliers did not change effects, they were included within the final analysis. Inspection of histograms, boxplots, and the skewness statistic (Shipiro-Wilk's test) indicated the assumption of normality was met. Greenhouse-Geisser corrections were considered unnecessary as there were only 2 levels of each independent variable. Results did not differ by sex, enrolment in first year psychology, or between participants for whom English was or was not their first language.

Table 1:

Means and Standard Deviations for Cycle One

		Recall	Mean JOL	Resolution
		M(SD)	*M(SD)*	*M(SD)*
0-100% Scale				
	Remember	23.26 (20.86)	36.36 (20.74)	0.07 (0.11)
	Forget	26.60 (20.72)	39.18 (17.38)	0.09 (0.09)
	Total	24.97 (20.61)	37.80 (18.92)	0.08 (0.10)
Binary				
	Remember	23.65 (18.18)	48.70 (27.48)	0.08 (0.10)
	Forget	28.72 (23.50)	30.83 (21.30)	0.10 (0.12)
	Total	25.92 (20.67)	40.68 (26.23)	0.09 (0.11)
Total				
	Remember	23.48 (19.18)	43.30 (25.28)	0.08 (0.10)
	Forget	27.66 (21.92)	35.01 (19.70)	0.10 (0.10)
	Total	25.48 (20.53)	39.34 (23.09)	0.09 (0.10)

Note: M = Mean. *SD* = Standard deviation.

Table 2:

Means and Standard Deviations for Cycle Two

		Recall	Mean JOL	Resolution
		M(SD)	*M(SD)*	*M(SD)*
0-100% Scale				
	Remember	0.52(0.30)	41.18(25.06)	0.13(0.13)
	Forget	0.58(0.31)	43.76(23.32)	0.17(0.18)
	Total	0.55(0.30)	42.5(23.93)	0.15(0.15)
Binary				
	Remember	0.54(0.27)	51.3(27.28)	0.14(0.12)
	Forget	0.56(0.31)	42.80(31.84)	0.19(0.18)
	Total	0.55(0.29)	47.48(29.41)	0.16(0.15)
Total				
	Remember	0.53(0.28)	46.87(26.55)	0.14(0.12)
	Forget	0.57(0.30)	43.28(27.59)	0.18(0.18)
	Total	0.55(0.30)	45.15(26.96)	0.16(0.15)

Note: *M* = Mean. *SD* = Standard deviation.

Recall Performance

"Recall" refers to memory performance on the recall components of the study. Higher scores indicate a greater percentage of words were correctly remembered during the recall test (better memory performance). Recall was expected to increase between cycle one and two, as the testing effect suggests additional learning and tests of information should promote greater recall in the second cycle (Karpicke & Roediger, 2008; Dunlosky, Rawson, Marsh, Nathan, Willingham, 2013). A paired samples *t*-test was used to determine the general pattern of effect for recall within the present study. Recall was significantly higher in cycle two than cycle one, $t(91) = 20.10$, $p < .001$, $95\%CI_{difference}[26.41, 32.21]$. This represented a large effect, demonstrating participant's memory for stimuli improved

with additional learning, d = 1.16, 95%CI[0.89, 1.42]. This finding is consistent with previous research (Hanczakowski et al., 2013; Serra & England, 2012).

JOLs

JOL values indicate the mean JOLs for each condition. As forget and remember JOLs have been converted to the same scale, higher JOLs indicate better predicted memory performance. A paired samples *t*-test was used to investigate the overall differences in JOLs between cycles. JOLs were significantly higher in cycle two than cycle one, although this represented a small effect, $t(91) = 2.65$, $p = .009$, $95\%CI_{difference}[1.46, 10.18]$, $d = 0.23$, 95%CI[0.02, 0.44]. This finding indicates participants predicted they would have better memory performance in cycle two than cycle one. This is consistent with previous research by Koriat (1997) and Hanczakowski et al. (2013).

Additionally, independent sample *t*-tests with Bonferroni adjustment ($\alpha = .025$) were used to investigate whether initial JOLs differed between frames. These tests were used as an indication of whether frames may promote the adoption of different anchors. Initial JOLs did not differ by frame when measured by 0-100% scales, but did differ on binary tasks. There was a small and non-significant difference between initial remember ($M = 36.36$, $SD = 20.74$) and forget frame JOLs ($M = 39.18$, $SD = 17.38$) when measured on a 0-100% scale, $t(41) = -.485$, $p = .630$, $95\%CI_{difference}[-14.59, 8.94]$, $d = 0.15$, 95%CI[-0.45, 0.75]. However, initial JOLs were significantly lower within forget frames ($M = 30.83$, $SD = 21.30$) than remember frames ($M = 48.70$, $SD = 27.48$) when measured on a binary scale, $t(47) = 2.50$, $p = .016$, $95\%CI_{difference}[3.48, 32.26]$. This difference represented a moderate effect, $d = 0.72$, 95%CI[0.14, 1.30]. This may indicate framing promoted the

adoption of lower anchors under binary-forget condition than the binary-rememeber condition. Hence while initial 0-100% scale JOLs did not differ by frame, initial binary JOLs did. As later discussed, this conflicts with Serra and England (2012).

Calibration

Calibration and over/underconfidence (O/U statistic) are commonly calculated to investigate metacognitive accuracy and determine whether judgements were overconfident, underconfident, or accurate. However, such calculations were believed to be inappropriate for the current study due to potential issue with calculating these statistics for binary tasks. Hence the association between recall and JOLs was assessed using a mixed ANOVA, within which JOLs and recall were included as a within-subjects factor. This method has previously been used by Koriat (1997) and Hanczakowski et al. (2013).

A 2 × 2 × 2 × 2 mixed model Analysis of Variance (ANOVA) containing Frame (remember or forget), Scale Type (0-100% or binary), Test Cycle (cycle one and cycle two) and Measure (JOLs and recall) was conducted. This analysis was used to test whether the increase in JOLs from cycle one to cycle two corresponded to the increases in recall between cycles (calibration). There was a significant interaction between cycle and measure, $F(1, 88) = 122.69$, $p < .001$, $\eta_p^2 = .582$. The effect of cycle therefore differed between recall and JOLs, with a greater increase in recall than JOLs between cycles. A larger effect size for the increase in recall ($d = 1.16$) than JOLs ($d = 0.23$) between cycles also illustrates this difference.

This pattern did not differ by frame or scale, as indicated by non-significant three-way interactions between cycle, measure, and scale, $F(1, 88) = .71$, $p = .402$, $\eta_p^2 = .013$, and between cycle, measure, and frame, $F(1, 88) = 1.20$, $p = .275$, $\eta_p^2 =$

.008. The four-way interaction between cycle, measure, scale, and frame was also non-significant, $F(1, 88) = 2.93$, $p = .091$, $\eta_p^2 = .032$. This supports that while the increase in recall was greater than the increase in JOLs between cycle, this pattern did not vary by frame or scale. Hence calibration did not differ by condition. This pattern is illustrated within Figures 2 – 5.

Confidence

In addition to calibration, it is important to consider the relationship between JOLs and recall by identifying whether manipulations promoted overconfidence, underconfidence, or accurate predictions of memory. By comparing JOLs to recall for each cycle, it is possible to determine whether participants predictions of memory performance were significantly overconfident (JOL > recall) or underconfident (JOL < recall). Small and non-significant difference between JOLs and recall indicate accurate predictions of memory performance. Paired samples *t*-tests with Bonferroni adjustment ($\alpha = .006$) were used to examine the effect of frame separately for each scale and cycle.

Minimal differences between frames were observed when JOLs were measured on a 0-100% scale. Yet the UWP effect was replicated under both frames. For the scale-remember condition, JOLs were higher than recall in cycle one, $t(20) = 2.60$, $p = .017$, $95\%CI_{difference}[2.60, 23.62]$, but lower than recall in cycle two, $t(20) = -2.51$, $p = .021$, $95\%CI_{difference}[-19.51, -1.79]$. Although neither comparison was significant following the Bonferroni adjustment, differences did constitute a moderate and small effect respectively, $d = 0.62$, 95%CI[0.15, 1.08], and $d = 0.40$, 95%CI[-0.05, 0.82].

Within the scale-forget condition, JOLs were also higher than recall in cycle one, $t(21) = 2.37$, $p = .027$, 95%$CI_{difference}$[1.55, 23.63], but lower than recall in cycle two, $t(21) = -2.72$, $p = .013$, 95%$CI_{difference}$[-24.24, -3.24]. Again, neither comparison was significant following the Bonferroni adjustment but each constituted a moderate effect, $d = 0.65$, 95%CI[0.18, 1.12], and $d = 0.50$, 95%CI[0.05, 0.94].[1] As can be observed in Figures 2 and 3, the UWP effect was replicated on both frames within 0-100% conditions. That is, JOLs were higher than recall (overconfident) in cycle one, but lower than recall (underconfident) in cycle two. Results on 0-100% scales therefore support the presence of the UWP effect, although this effect did not differ by frame.

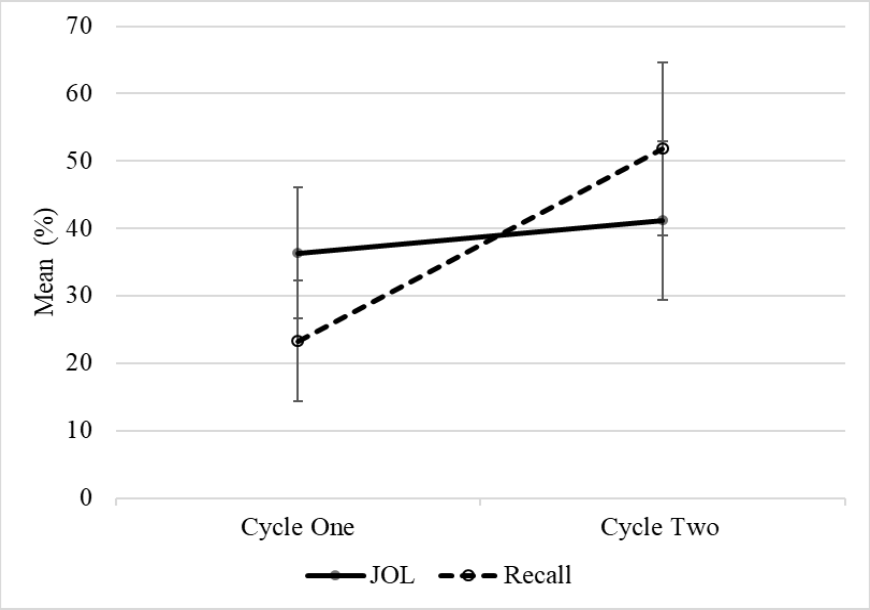


Figure 2: Remember frames when measured on a 0-100% scale. Error bars represent 95% confidence intervals

[1] Bonferroni adjustments have been criticised for being too conservative (Field, 2009). In the current example, moderate effects were considered a better indication of differences than non-significant *p* values following this correction.

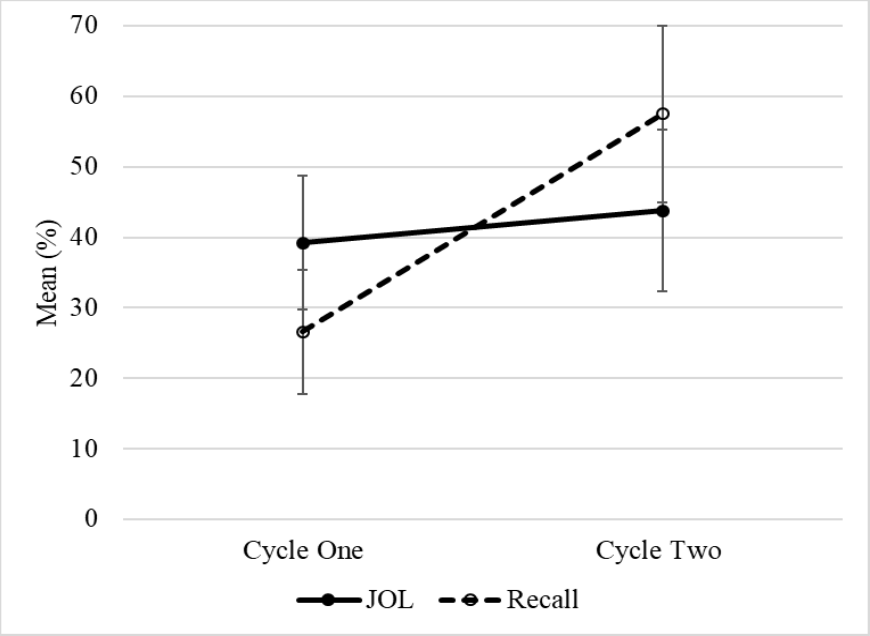


Figure 3: Forget frames when measured on a 0-100% scale. Error bars represent 95% confidence intervals.

Within binary conditions, the effect of framing differed depending on the test cycle. For the binary-remember condition, JOLs were significantly higher than recall in cycle one, $t(26) = 4.54$, $p < .001$, $95\%CI_{difference}[13.72, 36.40]$. This difference constituted a large effect, revealing participants in the remember condition were overconfident for cycle one, $d = 1.08$, 95%CI[0.60, 1.60]. A small and non-significant difference between JOLs and recall in cycle two suggests judgements were relatively accurate for this cycle, $t(26) = -0.47$, $p = .631$, $95\%CI_{difference}[-12.27, 7.57]$, $d = 0.22$, 95%CI[-0.16, 0.60].

For the binary-forget condition, JOLs were accurate in cycle one but underconfident in cycle two. A very small and non-significant difference between JOL and recall suggests JOLs were accurate for cycle one, $t(21) = 4.54$, $p = .634$, $95\%CI_{difference}[-7.02, 11.26]$, $d = 0.09$, 95%CI[-0.33, 0.51]. A moderate and significant difference between JOLs and recall suggests underconfidence in cycle

two JOLs, $t(21) = -4.25$, $p < .001$, $95\%CI_{difference}[-20.08, -6.89]$, $d = 0.43$, 95%CI[-0.01, 0.86]. Differences between frames on binary tasks can be observed in Figures 4 and 5.

Overall, the UWP effect was observed on 0-100% scale JOLs with minimal differences between frames. Binary-remember JOLs were overconfident on the first cycle, but became accurate. Binary-forget JOLs were accurate in the first cycle, but became underconfident. Hence the UWP effect was not found in the binary-remember condition, but a trend towards UWP was found in the binary-forget condition. Hence framing influenced JOL accuracy, but only when judgements were made using a binary task. While the increase in recall across cycles was greater than the increase in JOLs for all conditions, results suggest framing does affect JOLs. Yet as discussed later, these differences were not as predicted in the hypotheses.

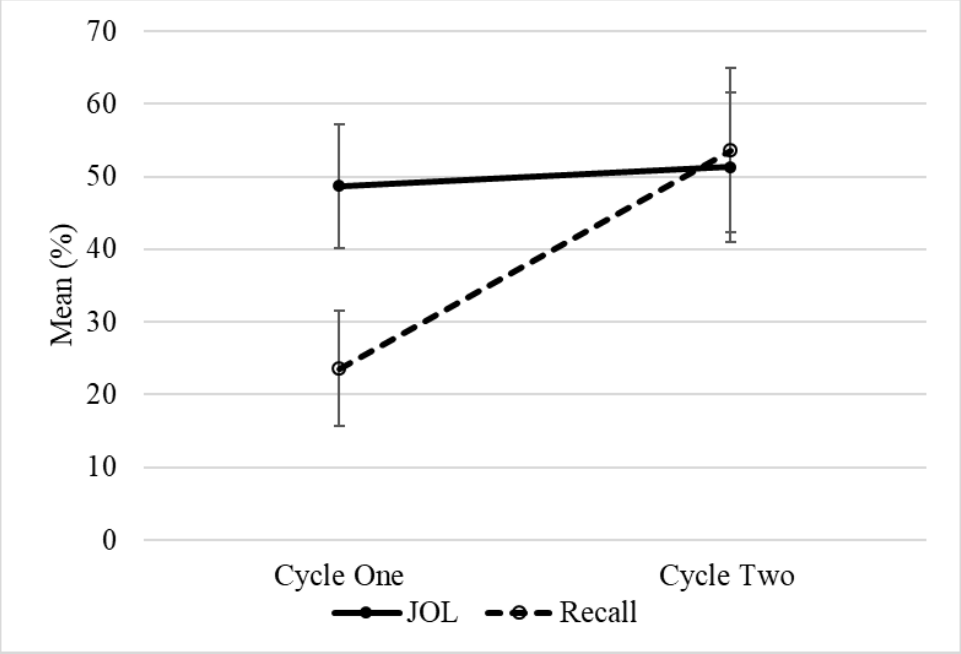


Figure 4: Calibration of the binary-remember condition. JOLs were overconfident in cycle one, but accurate in cycle two. Error bars represent 95% confidence intervals.

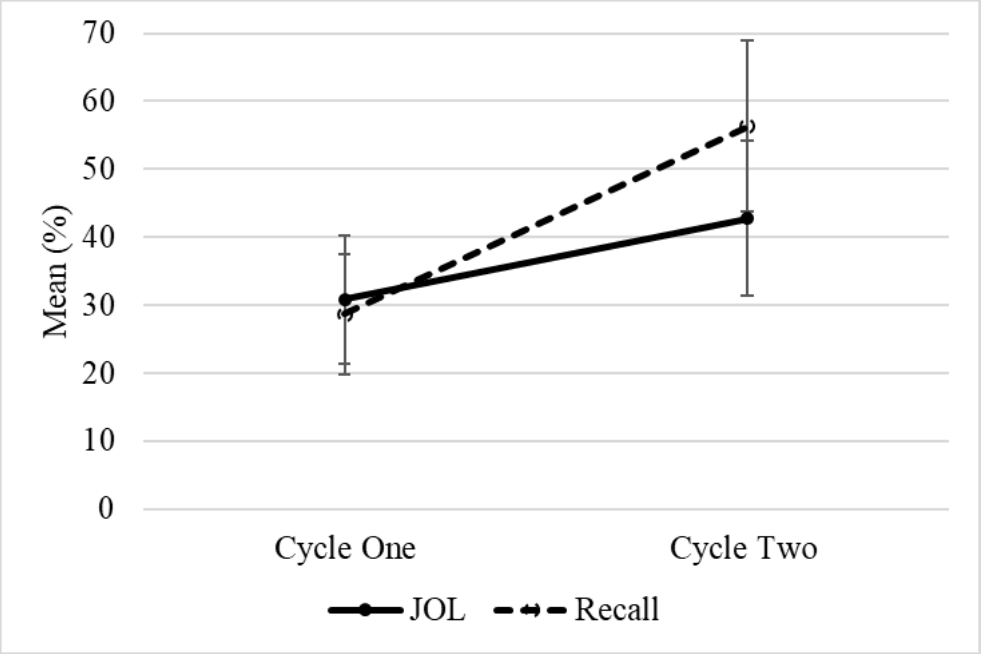


Figure 5: Calibration of the binary-forget condition. JOLs were accurate in cycle one, but underconfident in cycle two. Error bars represent 95% confidence intervals.

Resolution

A final analysis was conducted to investigate participants' ability to discriminate between items they would and would not recall. This ability – resolution – was analysed by comparing Adjusted Normalised Discrimination Index (ANDI) scores (Yaniv, Yates, & Smith, 1991). This statistic represents the ability to correctly discriminate between items which will or will not be recalled. ANDI scores range from 0 – 1, with higher scores indicating greater discriminability. Resolution was found to differ between cycles, but was minimally affected by manipulations of frame or scale. Resolution was greater in the second cycle than the first, as indicated by a significant main effect of cycle, $F(1, 84) = 16.81$, $p < .001$, $\eta_p^2 = .167$. The ability to discriminate between items that would and would not be recalled therefore

improved following additional exposure to information. This is constant with previous research (Hanczakowski et al., 2013).

Resolution did not differ as the result of any of the other manipulations. There were non-significant interactions between frame and cycle, $F(1, 84) = .61$, $p = .439$, $\eta_p^2 = .007$ and between scale and cycle, $F(1, 84) = .004$, $p = .951$, $\eta_p^2 < .001$. There was a non-significant interaction between frame, scale, and cycle, $F(1, 84) = .00$, $p = .996$, $\eta_p^2 < .001$. Hence resolution differed between cycles, but was minimally affected by manipulations of frame or scale.

Discussion

The present study investigated the impact of manipulating measurement variables on metacognitive accuracy. Of greatest interest was whether JOL question frame would influence metacognitive accuracy. Based on previous research, three results from manipulating framing were considered plausible. It was predicted forget frames may produce greater JOL accuracy on both scales, that framing effects could be present on 0-100% scales but minimally affect binary tasks, or that framing would minimally affect JOL accuracy. None of these predictions were completely supported within the current study. Converging evidence for the effect of framing on multiple scales was not obtained, as binary measures may be more sensitive to the effects of framing. Responses on 0-100% scales most resemble Rhodes and Castel (2008) as there was minimal effect of framing. However, the effect of framing within binary conditions was not consistent with any of the previous framing literature reviewed. Hence a fourth possible pattern for the effect of framing was observed. Explanations for the effect of framing most resemble those of Serra and England (2012), as use of different anchoring may explain differences between frames. Yet

the relation between these anchors may be the opposite of what Serra and England predicted, as forget frame JOLs may be anchored lower than remember frame JOLs.

Effect of Manipulations on Recall Performance and JOLs

Recall improved from cycle one to cycle two. This suggests learning occurred between the first and second recall tests, and confirms participants were engaging in the task. This finding is consistent with the testing effect (Karpicke & Roediger, 2008) and previous metacognitive literature which has shown improved recall in later cycles following cued recall of word-pairs (Hanczakowski et al., 2013; Serra & England, 2012). It also conforms with learning literature, which demonstrates restudy and repeated retrieval of information improves memory strength and facilitates recall of information (Dunlosky et al., 2013; Wiklund-Hörnqvist et al., 2014). Neither framing nor scale type affected recall. This was expected, as these manipulations have been demonstrated to affect JOLs rather than memory performance (Finn, 2008; Hanczakowski et al., 2013). This finding is therefore consistent with previous literature (Finn, 2008; Hanczakowski et al., 2013).

In addition to greater recall, higher JOLs in cycle two indicates participants believed memory performance would be better in cycle two than cycle one. While JOLs differed significantly between cycles, this difference constituted only a small effect. This is in line with previous research, although a weaker effect (Hanczakowski et al., 2013). Overall results for recall and JOLs were therefore as expected.

Resolution

Resolution (relative accuracy) was greater in cycle two than cycle one. This suggests the ability to discriminate between items that would or would not be recalled improved following additional learning. Differences in resolution between cycle one and two are consistent with Hanczakowski et al. (2013), who found resolution increased across cycles on both scales. This finding is also consistent with research on the MPT heuristic, which suggests using information about previous memory performance can improve the ability to predict memory for the same information during subsequent tests (Finn & Metcalfe, 2008; Serra & Ariel, 2014). Resolution therefore benefited from additional study and testing, conforming with previous research (Hanczakowski et al., 2013; Koriat, 1997).

Significant differences in resolution were not observed following the manipulation of frame. This suggests framing does not change the relative accuracy of participant's metacognitive decisions. For cycle one, this is consistent with Finn (2008), and Rhodes and Castel (2008), who found minimal differences in resolution between frames. However, when compared to multi-cycle studies such as Serra and England (2012), present results conflict somewhat. As within the current study, Serra and England found equivalent resolution under both frames during the first cycle. Yet unlike the present study, Serra and England found remember frames promoted greater resolution than forget frames in cycle two. Following their results, Serra and England proposed forget frames impaired use of the MPT heuristic.

Current results for resolution in cycle two conflict with this notion. Minimal differences in resolution between frames in cycle two may suggest framing does not affect use of the MPT heuristic. It is possible framing does not alter resolution, or provide cues towards discriminability. Alternatively, frames may provide cues about

resolution but participants did not rely upon them. Following the suggestion of Finn (2008), reliance upon cues believed to be most predictive of memory performance may reduce the use of other cues when making metacognitive decisions. Participants in the current study may therefore have believed the MPT heuristic was a better indication of memory performance than cues provided by framing. Hence framing may not influence the resolution of subsequent metacognitive decisions, while the MPT heuristic may explain why resolution improves between cycles.

0-100% Scale Conditions

JOLs made on 0-100% scales were minimally influenced by framing, but did display the UWP effect. The effect of framing on 0-100% scale JOLs replicated the findings of Rhodes and Castel (2008). As within their study, framing had a minimal effect on metacognitive accuracy when a 0-100% scale was used. Regardless of frame, JOLs made on 0-100% scales replicated the UWP effect. JOLs were overconfident on initial testing, yet underconfident during the second cycle for both frames. Hence findings on 0-100% scales within the current study support the prediction there may be minimal effect of framing (Rhodes & Castel, 2008) and support hypotheses about the UWP effect (Koriat, Sheffer, & Ma'ayan, 2002).

Minimal differences between frames when JOLs were measured using 0-100% scales may have resulted from how these judgements were made. Serra and England (2012) suggested framing may result from anchoring effects, with forget frames anchored at the midpoint and remember frames anchored between 20-40%. It is possible the way JOLs were made on 0-100% scales within the current study disrupted the formation of these frame related anchors. On 0-100% scales, a marker on a bar from 0 to 100 was moved to rate JOLs. This method was considered

beneficial to reduce typing errors. The method was also considered more sensitive, as subtle differences in responses may be missed if a scale with 10% increments was used. Yet it may have had unintended consequences on anchoring effects. As the marker always appeared in the middle of the bar (50%), it perhaps prompted participants to adopt an anchor at the same value. This visual cue may have resulted in participants considering this anchor as more salient or diagnostic than the anchors provided by framing effects. Hence the method of making 0-100% scale JOLs may have minimised the effects of framing.

To address this, future studies could replicate the method of the current experiment but require typed responses on 0-100% scales. However, it would be best to ensure only numbers between 0 and 100 were accepted. However, this method may also be more susceptible to participants restricting responses to increments of five or ten, or refraining from providing ratings at the extremes of the scale (Mickes, Hwe, Wais, & Wixted, 2011; Mickes, Wixted, & Wais, 2007). Hence researchers should be aware this method may result in participants failing the use the full range of the scale. Alternatively, it may be possible for the response bar to not have a marker, and the marker to only appear after the participant clicked to make their JOL. Minimal differences between frames following this correction would provide additional evidence for Rhodes and Castel's (2008) finding framing minimally affects JOLs. Alternatively, it is possible this change will JOLs will differ between frames. Depending on the pattern of effects, this could be consistent with the binary condition results of the present study, or support Serra and England's (2012) research. Changing how responses are provided on 0-100% scales would therefore allow investigation of whether methodological differences between scales produced results.

Binary Task Conditions

If multiple scales had not been included, it may have been concluded framing does not affect metacognitive accuracy. Instead, results from the binary conditions suggest framing is having an effect. JOLs made in the binary-remember condition displayed overconfidence in memory during the first cycle, but accurate predictions during the second cycle. Unlike remember frames, binary-forget JOLs displayed accurate predictions within the first cycle but a trend towards underconfidence within the second. Interestingly, binary results within the current study are almost the opposite of Serra and England's (2012) findings on 0-100% scales (Serra and England found remember frames were well calibrated but became underconfident, while forget frames were overconfident but became well calibrated). Overall, while differences between frames were observed, results were not consistent with any of the predicted patterns. Hence this study conflicts with hypotheses about framing and provides a fourth pattern of effects.

Findings on binary measures have some implications for understanding the effect of scale type on the UWP effect (Hanczakowski et al., 2013; Koriat, Sheffer, & Ma'ayan, 2002). Consistent with Hanczakowski et al.'s (2013) results, the UWP effect was found on 0-100% scales but not on binary tasks. However, this pattern was only observable under a remember frame within the current study. Instead, a trend towards UWP within the binary-forget condition was observed within current results. This suggests an important boundary condition to Hanczakowski et al.'s findings; while the UWP effect may be eliminated by binary JOLs, this only applies if JOLs are made in a remember frame. What remains clear is differences between scales exist, and can influence the accuracy of some metacognitive decisions. Results therefore support Hanczakowski et al. and differing interpretations of the JOL

question depending on scale type, but only when considering judgements made under a remember frame.

One explanation of findings from binary tasks is that framing affects how difficult it is to make accurate metacognitive decisions. This idea resulted from literature on retrospective memory and eyewitness identifications. Retrospective memory is memory for information, events, or experiences encountered in the past (Kvavilashvili & Ford, 2014). Research in this area suggests people are better at determining information has been seen before (remember something has been encountered) than determining information has not been encountered (Weber & Brewer, 2004, 2006). Underpinning this finding is the idea it may be easier to collect memorial evidence a stimulus has been encountered than it is to determine a stimulus does not matches the memory of anything previously encountered (Weber & Brewer, 2004, 2006).

Following these ideas, the framing effects observed within binary conditions could potentially be explained by differences in the difficulty of the metacognitive task. Overconfidence in memory may be produced if people find it easier to make accurate metacognitive decisions when asked to assess the likelihood of remembering. In comparison, if assessing the likelihood of forgetting is more difficult, it is possible this will lower confidence in memory or produce underconfidence. High confidence in the binary-remember condition, but underconfidence in the binary-forget condition could therefore be explained by differences in the difficulty of making metacognitive assessments.

Despite the intuitive relevance of theories relating to retrospective memory, it may not be applicable to the current study. These theories do not explain why framing did not affect both scales equally. Should assessing forgetting be more

difficult than assessing remembering, underconfidence should have been observed within the forget conditions of both scales. Additionally, judging whether something was encountered previously may rely upon different processes than predicting the likelihood of future recall. Theories relating to how difficult it is to assess memorial evidence may therefore be less relevant for judging future recall. Hence alternative explanations of findings are considered more plausible.

A more plausible explanation of binary task results could be the interaction of two factors; differences in anchoring and greater sensitivity of binary measures to framing effects. First, anchoring effects may underpin why there were differences in JOLs between framing conditions. Serra and England (2012) suggested the differences they found between frames resulted from use of different anchors. Based on their results, it was suggested forget frames were anchored higher (around 50%) than remember frames (anchored between 20 – 40%). However, results from the binary conditions of the current study suggest the opposite. As can be observed in cycle one of Figures 4 and 5, initial forget frame JOLs (M = 30.83) were significantly lower than initial remember frame JOLs (M = 48.70). Hence while SOJ were not conducted to identify the value of each anchor, initial JOLs on the binary scale suggest forget frames may be anchored comparatively lower than remember frames.

Use of different anchors could explain the observed binary results and provide a potential explanation of framing effects. Lower anchors for forget frames may have resulted in participants displaying reasonably accurate metacognitive judgements in the first cycle on binary scales. As anchoring results in minimal JOL change between cycles (Scheck & Nelson, 2005), lower anchors for forget frames compared to remember frames would explain why underconfidence was found in the

second cycle of the binary-forget condition. In comparison, higher anchors in the binary-remember frame may have promoted the overconfidence in memory observed within the first cycle. In the second cycle, well-calibrated judgements may have been due to memory accuracy coincidentally improving to the same level as the unchanged JOLs. Hence differences between frames may have been caused by the adoption of different anchors.

A second factor which could have contributed to results is differences caused by scales. This factor may explain why differing patterns of framing were observed between scales. Based on Hanczakowski et al.'s (2013) research, binary tasks may assess the probability of recall, while 0-100% scales assess confidence in recall. In addition to this, Finn (2008) suggests forget frames may promote JOL accuracy by increasing reliance on predictive memory cues. However, it should be noted minimal differences between forget and remember frames on 0-100% scales may indicate this idea is more fragile than Finn predicted. Following these two arguments, it is suggested framing effects on binary but not 0-100% scales indicates framing has greater influence on the probability of recall than confidence in recall. It is therefore possible binary scales may be more sensitive to the effects of framing than 0-100% scales. Differing effects of framing depending on scale type could therefore be explained by use of different anchors between frames, and binary tasks being more sensitive to framing.

Implications

The study indicates simply getting students or clients to respond to a yes-no question about whether information is likely to be forgotten could promote accurate ratings of predicted memory performance. This may have practical implications for

health professionals, educators, or other professions where it is important clients are able to accurately recall information. Within all these settings, accurate metacognitive judgements are required to determine what information is known, and what may require future learning. Without a method to accurately assess this, it becomes difficult to identify when it is necessary to seek further information (Medina, Castleberry, & Persky, 2017). For example, people who are unable to accurately determine further learning is required may not gain all necessary health or drug information from their doctors. The current research suggests binary tasks under a forget frame may be a simple method to help people identify whether information will be recalled. While confirmation of findings is needed, binary-forget tasks may be a practical method for improving learning efficacy in health scenarios, education settings, and other real-world settings.

An additional implication of the current study is it highlights the need for further research into the effects of manipulating frames. Overall framing results differed distinctly from all previous framing research addressed within this paper. A fourth possible result of manipulating JOL question frame has therefore been observed, meaning conflicting results between framing studies remain. Hence the study highlights more research is needed to establish a pattern of effects of framing. Such research could follow the method of Serra and England's (2012) paper, as collecting both JOL and SOJ could provide deeper understanding of anchoring effects. Inclusion of both binary tasks and 0-100% scales would be beneficial, as the present study highlights the effect of framing can differ depending on the measure used. Doing so may provide additional evidence that framing does not affect anchoring. Alternatively, such research may replicate the pattern of anchoring observed within either the present study or Serra and England's study. Such research

would be beneficial as it could improve understanding of the impact of manipulating framing.

The present study adds three ideas to current understanding of the effects of measurement variables on JOL accuracy. First, the study builds on ideas such as Prospect Theory by suggesting the way JOL assessments are framed can also influence the answers participants provide. Second, the effects of framing may result from forget frames promoting comparatively lower anchors than remember frames. Finally, the study adds to research on differences between scales by indicating binary tasks may be more sensitive to framing than 0-100% scales. Collectively, these ideas indicate the accuracy of metacognitive assessments such as JOLs can be influenced by how they are measured. These three findings add to the current understanding of metacognition, and indicate methods of metacognitive assessment should be carefully designed to promote accurate judgments and decrease cognitive biases.

A final implication of the present study relates to how future research may be conducted. Following recommendations from Hanczakowski et al. (2013), the study provides evidence that examining the results of manipulations on multiple scales can improve the understanding of effects. In relation to framing, no effect of framing was found on 0-100% scales. As previous framing literature has only considered 0-100% scales, the present study adds information about how the result of manipulating framing changes depending on scale type. The disparity between framing effects on different scales highlights it may be premature to make conclusions about a manipulation without considering converging evidence from multiple scales. The present study therefore highlights that inclusion of multiple scales could be a valuable tool for gathering converging evidence and investigating the true effect of manipulations on metacognitive accuracy. It may therefore have implications for

future research. This research could assess whether manipulating the difficulty of the recall task (Scheck & Nelson, 2005) or delaying JOLs (van Loon et al., 2013) provides similar results on binary tasks to what has previously been observed on 0-100% scales. Such research would provide v information on the implications of and mechanisms behind such manipulations.

Limitations

A limitation of the current and other framing studies is how remember and forget frames were compared. Within the present study, forget frames were converted to the same scale as remember frames by reverse scoring forgetting. This transformation was believed necessary to directly compare frames. The same method was used within previous framing studies (Finn, 2008; Rhodes & Castel, 2008; Serra & England, 2012). However, forgetting and remembering may not be direct opposites (Roediger, Dudai, & Fitzpatrick, 2007). Instead, it is possible each judgment may require different information, or represent the assessment of different outcomes (Roediger et al., 2007). Conducting this transformation therefore ignores the possibility forgetting and remembering may not be fundamentally opposites (i.e. that forgetting may not be equal to 1 – the likelihood of remembering). As such, it may be beneficial to consider the effects of each frame separately rather than converting frames to the same scale. Future framing research may therefore benefit from an alternate method of comparing frames.

A second potential limitation of the present study relates to poor recall performance. Low recall scores could potentially indicate the recall task is too complicated, perhaps making it too difficult for participants to differentiate between what would and would not be recalled. Poor resolution in cycle one (M = 0.09) and

cycle two (*M* = 0.16), and poor recall performance (*M* = 25.48) in cycle one may support this notion. However, low recall scores within the first cycle are common within metacognition literature. Using the same word list and similar word-pair presentation time (4 seconds) as the present study, Serra and England (2012, experiment 1) found recall was under 30% for both frames during the first cycle. Likewise, Finn (2008) found mean recall was 17% (*SE* = 2) under a remember frame and 19% (*SE* = 3) for forget frames when completing a task akin to the current study. Poor recall in the first cycle of the present study (25.47%) is therefore comparable to results on similar metacognitive tasks. In addition, both resolution and recall performance did improve between cycles. This indicates the task was not so difficult it prevented learning or improvements to the ability in resolution. Hence poor recall in the first cycle should not be considered a limitation.

A final limitation is it remains unclear whether the method of assessing JOLs promoted differences in metacognitive accuracy, or simply altered the ability to translate beliefs about recall likelihood onto an external scale. Based on the current study, it is possible differences may have greater relation to changes in reporting than changes to metacognitive accuracy. Yet this idea was not explicitly explored in the current study, and further investigation into this idea is required. Regardless, results do suggest the method of assessment affects the JOL ratings given. Further investigation of why this occurs would be beneficial to clarify this distinction.

Concluding Comments

The present study provides support for the notion both the way metacognitive questions are framed, and the scale used to measure such judgements can affect the accuracy of immediate JOL decisions. Despite Hanczakowski et al.'s (2013) finding

the UWP effect occur on 0-100% scales but not binary measures, the current research was only able to replicate this pattern within remember frames. The theory scales promote either a probability or confidence interpretation of JOL questions may therefore be less robust than originally predicted. Results were not consistent with any of the previous framing studies addressed, hence highlighting the need for more research to clarify conflicting findings. Framing within the current study could be explained by lower anchors within forget than remember frames, and greater sensitively of binary tasks to the effects of framing.

While methodological differences between scales was considered a less plausible explanation of results, it cannot be completely dismissed. The study may also be limited by how remember and forget frames were compared. Future studies could correct these issue by changing the methods of assessing JOLs on 0-100% scales, and developing an alternate method of comparing frames. The study has theoretical implications for research into both framing and scale effects on immediate JOLs, and highlights the need to establish a pattern of effects in both areas. Most importantly, the study reveals the recorded accuracy of metacognitive decisions can depend on the measure used. Research into and development of methods to accurately assess metacognition in both real-word and research settings is therefore necessary.

References

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences.* New Jersy: Lawrence Erlbaum Associates.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Dunlosky J., Rawson K. A., Marsh E. J., Nathan M. J., Willingham D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. doi:10.1177/1529100612453266

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. a. (2005). Second-order judgments about judgments of learning. *The Journal of General Psychology, 132*(4), 335–346. doi: 10.3200/GENP.132.4.335-346

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. Retrieved from: https://link.springer.com/content/pdf/10.3758%2FBF03193146.pdf

Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). London: Sage Publications Ltd.

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36*(4), 813–821. doi: 10.3758/MC.36.4.813

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*(1), 19–34. doi: 10.1016/j.jml.2007.03.006

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice

effect. *Journal of Memory and Language, 69*(3), 429–444. doi:10.1016/j.jml.2013.05.003

Jonsson, A. C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgements over time, content domain, and gender. *Personality and Individual Differences, 34*(4), 559–574. doi:10.1016/S0191-8869(02)00028-4

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292. Retrieved from http://www.jstor.org/stable/1914185

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 15*(319), 966–968. doi: 10.1126/science.1152408

Kelly, D., & Donaldson, D. (2016). Investigating the complexities of academic success: Personality constrains the effects of metacognition. *The Psychology of Education Review, 40*(2), 17–24. Retrieved from http://dspace.stir.ac.uk/handle/1893/24578

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology,126*(4), 349–370. doi: 10.1037/0096-3445.126.4.349

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology, 131*(2), 147–162. doi: 10.1037/0096-3445.131.2.147

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology, 138*(4), 449–468. doi: 10.1037/a0017350

Kvavilashvili, L., & Ford, R. M. (2014). Metamemory prediction accuracy for

simple prospective and retrospective memory tasks in 5-year-old children. *Journal of Experimental Child Psychology, 127*, 65–81. doi: 10.1016/j.jecp.2014.01.014

Medina, M. S., Castleberry, A. N., & Persky, A. M. (2017). Strategies for improving learner metacognition in health professional education. *American Journal of Pharacetical Education, 81*(4). doi: 10.5688/ajpe81478

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174–179. doi: 10.3758/PBR.15.1.174

Metcalfe, J., & Kornell, N. (2005). A Region of Proximal Learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463–477. doi: 10.1016/j.jml.2004.12.001

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology, 140*(2), 239–257. doi: 10.1037/a0023007

Mickes, L., Wixted, J., & Wais, P. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14*(5), 858–865. doi: 10.3758/BF03194112

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The delayed-JOL effect. *Psychological Science, 2*, 267-270. doi: 10.1111/j.1467-9280.1991.tb00147.x

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology, 137*(4), 615–625. doi: 10.1037/a0013684.

Roediger, H. L., Dudai, Y., & Fitzpatrick, S. M. (2007). Science of memory:

Concepts. Oxford: Oxford University Press.

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect : Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology, 134*(1), 124–128. doi: 10.1037/0096-3445.134.1.124

Schneider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring. *Cognitive Development, 15*(2), 115–134. doi: 10.1016/S0885-2014(00)00024-1

Schmitz, C. (2015). LimeSurvey: An open source survey tool (Version 2.06). Hamburg, Germany: LimeSurvey Project.

Serra, M. J., & Ariel, R. (2014). People use the memory for past-test heuristic as an explicit cue for judgments of learning. *Memory and Cognition, 42*(8), 1260–1272. doi: 10.3758/s13421-014-0431-0

Serra, M. J., Dunlosky, J., & Hertzog, C. (2008). Do older adults show less confidence in their monitoring of learning? *Experimental Aging Research, 34*(4), 379–391. doi: 10.1080/03610730802271898

Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgements of remembering and forgetting. *The Quarterly Journal of Experimental Psychology, 65*(11), 2231–2257. doi: 10.1080/17470218.2012.685081

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science, 22*(11), 1359–1366. doi: 10.1177/0956797611417632

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Allyn & Bacon.

Undorf, M., Böhm, S., & Cüpper, L. (2015). Do judgments of learning predict

automatic influences of memory? *Journal of Experimental Psychology, 42*(6), 882–896. doi: 10.1037/xlm0000207

van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merrienboer, J. J. G. (2013). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning, 8*(2), 173–191. doi: 10.1007/s11409-013-9100-0

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology, 10*(3), 156–172. doi: 10.1037/1076-898X.10.3.156

Weber, N., & Brewer, N. (2006). Positive versus negative face recognition decisions: Confidence, accuracy, and response latency. *Applied Cognitive Psychology, 20*(1), 17–31. doi: 10.1002/acp.1166

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology, 55*(1), 10–16. doi: 10.1111/sjop.12093

Xu, J., & Metcalfe, J. (2016). Studying in the region of proximal learning reduces mind wandering. *Memory & Cognition, 44*(5), 1–15. doi: 10.3758/s13421-016-0589-8

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*(3), 611–617. doi: 10.1037/0033-2909.110.3.611

Zawadzka, K., & Higham, P. A. (2015). Judgments of learning index relative confidence, not subjective probability. *Memory and Cognition, 43*(8), 1168–1179. doi:10.3758/s13421-015-0532-4

Appendices

Appendix A: Ethics Approval

Email received from Katherine Shaw (katherine.shaw@utas.edu.au) on Wednesday 15th March 2017 at 2:11 pm.

Ethics Ref No: H0012660
Project title: Confidence in memory

This email is to confirm that the following amendment was approved by the Chair of the Tasmania Social Sciences Human Research Ethics Committee on 15/3/2017:

- Addition of student researchers Mr Rod Garton, Ms Amelia Kohl, Ms Morgan Norris, Mr Rafal Kozlowski, Ms Talira Kucina and Ms Rachel Breen.
- Removal of student researchers Ms Rebecca Healy, Ms Kate Edwards, Ms Catherine Bishop, Ms Katie-Lee Crawford, Ms Katie Henderson, Ms Rebecca Kaiser, Mr Robert Kirkis, Mr Michael O'Leary and Mr Tane Thomas.

All committees operating under the Human Research Ethics Committee (Tasmania) Network are registered and required to comply with the National Statement on Ethical Conduct in Human Research (NHMRC 2007, updated May 2015).

This email constitutes official approval. If your circumstances require a formal letter of amendment approval, please let us know.

Should you have any queries please do not hesitate to contact me.

Kind regards
Katherine

Katherine Shaw
Executive Officer, Social Sciences HREC
Office of Research Services | Research Division
University of Tasmania
Private Bag 1
Hobart TAS 7001
T +61 3 6226 2763
[www.utas.edu.au/research]www.utas.edu.au/research

UNIVERSITY of TASMANIA

CRICOS 00586B

Appendix B: Participant Information and Consent Sheet

You are invited to participate in a study investigating metacognition. Metacognition is our ability to think about our thinking. This includes making judgments about how well we have learnt information, and our confidence in our memory.

This study is being conducted by Rachel Breen as part of an Honours Degree at the University of Tasmania. Matthew Palmer is supervising the study. To speak with the researchers please contact either Rachel Breen (rjbreen@utas.edu.au; Ph: 03 6324 3004) or Matthew Palmer (matt.palmer@utas.edu.au; Ph: 03 6324 3004).

1. What is the purpose of this study?

The purpose of this study is to investigate how well people make metacognitive decisions. This study will help inform researchers about how these decisions are made, and will help inform future research.

2. Why have I been invited to participate?

We are seeking volunteers to participate in the study. Anyone over the age of 18 with normal or corrected to normal vision can participate.

3. What will I be asked to do?

You will be asked to complete a series of tasks which involve learning and being tested on a list of word pairs. As part of this, you will be asked about your confidence in your memory. Tasks will be completed individually on a provided computer. The study will take approximately 60 minutes to complete.

4. Are there any possible benefits from participation in this study?

Participating in this study is a great opportunity to see how research works. In addition, your participation will help researchers gain a better understanding of

memory and metacognition. This information can help guide future research and investigate ways to help people learn more effectively.

You will also be offered either financial compensation or course credit in return for your time. First year psychology students will be offered either 60 minutes course credit or $20. Those not enrolled in first year psychology will be offered $20 as a thank you for your time.

5. Are there any possible risks from participation in this study?

There are no anticipated risks from participating in this study. The potential for the experiment to cause harm is expected to be minimal. Should support be required following the study, free counselling services can be contacted on the Mental Health Hotline: 1800 332 388 (Tasmania) or Lifeline: 13 11 14 (National). Current students can also access free counselling services at the University of Tasmania (Launceston: 03 6324 3787; Cradle Coast: 03 6430 4947).

6. What if I change my mind during or after the study?

You are free to withdraw from the study at any time without explanation. There will be no negative consequences if you decide to leave.

7. What will happen to the information when this study is over?

All data will be anonymous. No identifying information will be collected from participants. Data will be kept on a secure computer at the University of Tasmania. Only the researchers will have access to it. Following the completion of the study, data will be kept for a period of five years. After this time, all data will be archived.

8. How will the results of the study be published?

The results of this study will be presented verbally at a presentation day and in the honours thesis as per the requirements of the Honours degree. Data will be de-identified; you will not be able to be identified based on data presented. A summary

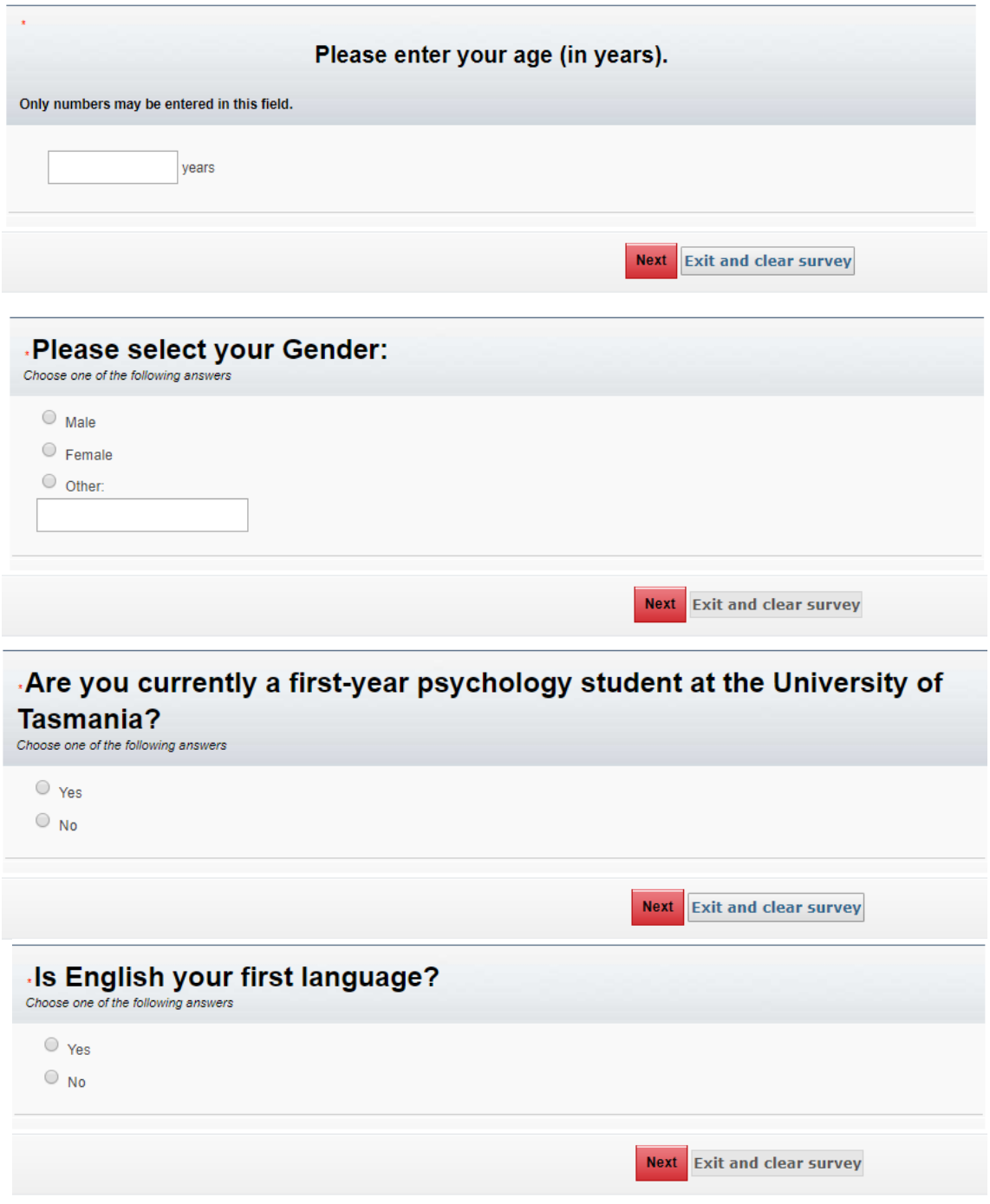
of results will be available on the University of Tasmania's website (http://www.utas.edu.au/psychology/research/research-project-reports) following the completion of the study.

9. What if I have questions about this study?

If at any point you wish to speak with the researchers, you may contact either Rachel Breen (rjbreen@utas.edu.au; Ph: 03 6324 3004) or Matthew Palmer (matt.palmer@utas.edu.au; Ph: 03 6324 3004). This study has been approved by the Tasmanian Social Sciences Human Research Ethics Committee. If you have concerns or complaints about the conduct of this study, please contact the Executive Officer of the HREC (Tasmania) Network by phone (+61 3 6226 6254) or email (human.ethics@utas.edu.au). The Executive Officer is the person nominated to receive complaints from research participants. Please quote the ethics reference number H0012660.

□ I have read and understood the information provided about this study, and I consent to participating. *[Ticking this box was required to move to the next page and begin the study]*.

Appendix C: Demographic Questions



Appendix D: Word Pairs

BufferA	Arm - Market	PAIR17	Woods - Chin
BufferB	Icebox - Acrobat	PAIR18	Church - Mammal
BufferC	Banner - Nun	PAIR19	Claw - Salad
PAIR01	Macaroni - Bar	PAIR20	Jail - Coffee
PAIR02	Barrel - Star	PAIR21	Glacier - Cord
PAIR03	Beast - Fabric	PAIR22	Corn - Planet
PAIR04	Vest - Bird	PAIR23	Cotton - Reptile
PAIR05	Blister - Cabin	PAIR24	Diamond - Umbrella
PAIR06	Daffodil - Blood	PAIR25	Monarch - Doll
PAIR07	Blossom - Locker	PAIR26	Door - Officer
PAIR08	Rattle - Board	PAIR27	Slipper - Dove
PAIR09	Lawn - Book	PAIR28	Ticket - Dummy
PAIR10	Piston - Boulder	PAIR29	Sunburn - Elephant
PAIR11	Pelt - Brain	PAIR30	Flag - Window
PAIR12	Bronze - Whale	PAIR31	Flesh - Kettle
PAIR13	Bullet - Yacht	PAIR32	Foam - Meadow
PAIR14	Candy - Prairie	PAIR33	Suds - Fowl
PAIR15	Cat - Jury	PAIR34	Potato - Frog
PAIR16	Cellar - Elbow	PAIR35	Fur - Oats

PAIR36	Glass - Journal	PAIR50	Lump - Wine
PAIR37	Volcano - Hammer	PAIR51	Bowl - Missile
PAIR38	Lake - Harp	PAIR52	Mountain - Skin
PAIR39	Hillside - Revolver	PAIR53	String - Mule
PAIR40	Poet - Home	PAIR54	Oven - Ship
PAIR41	Hoof - Slave	PAIR55	Toy - Pencil
PAIR42	Hotel - Noose	PAIR56	Python - Building
PAIR43	Harness - Snake	WP57	Spinach - Typhoon
PAIR44	Ink - Lark	WP58	Sugar - Prison
PAIR45	Iron - Leopard	WP59	Corpse - Forest
PAIR46	Juggler - Mast	WP60	Fox - Pudding
PAIR47	Leaflet - Tower	BufferD	Butter - Hospital
PAIR48	Seat - Letter	BufferE	Person - Storm
PAIR49	Doctor - Lobster	BufferF	Fire - Apple

Appendix E: Task Instructions

WELCOME

[Demographics Questions Presented]

Before beginning this study, please make sure your environment is as distraction free as possible. For example, it would be best to either put your mobile phone on silent or turn it off to ensure it will not distract you. Many of the components of this study are timed. You will not be able to pause the study after it has begun.

In this study, you will be shown word-pairs containing two English words. For example:

photo - table

(Cue word) *(Target word)*

Your task is to learn these words to the best of your ability in preparation for a test. You will be given 3.5 seconds to study each word-pair.

During the test, you will be shown the first word in the pair, and will be asked to type the second word. After you study each word-pair, you will be asked to indicate whether you are likely to forget the second word when the first English word is presented in the upcoming test.

[For the binary-forget condition] You will be asked to respond by selecting either “Yes" or "No", whereby:

Yes means you are likely to forget,

And No means you are not likely to forget.

[For the binary-remember condition] You will be asked to respond by selecting either “Yes” or “No”, whereby:

Yes means you are likely to remember,

And No means you are not likely to remember.

[For the scale-remember condition]: You will be asked to rate how likely you are to remember on a scale from 0% - 100%, whereby:

0% means you are not likely to remember,

And 100% means you are very likely to remember.

[For the scale-forget condition]: You will be asked to rate how likely you are to forget on a scale from 0%-100%, whereby:

0% means you are not likely to forget,

And 100% means you are likely to forget.

When you are ready to begin the study phase, click next.

[Word-Pairs Cycle One Presented]

Now you have 2 minutes to solve as many of the following maths problems as possible. Please enter your answers in the correct order in the box below. Separate each answer with a comma (,).

[Filler Task One Presented]

You will now undergo a test to see how many of the word pairs you can remember. During the test, the first word will be displayed and you will be asked to type the second word of the pair. This part of the study is not timed, and you may take as long as you like. If you do not know the answer, you may type "X" to indicate this. Once you are satisfied with your answer, press NEXT to move to the next screen.

[Cued Recall Task Presented]

You now have a two-minute break before the beginning of the next part of the study. Please take this time to stretch and relax. The study will resume automatically after two minutes has elapsed.

You will now be given the opportunity to study the word-pairs again in preparation for a final test. You will be given 3.5 seconds to study each word-pair. During the test, you will be shown the first word in the pair, and will be asked to type the second word.

After you study each word-pair, you will be asked to rate how likely it is you will remember/forget the second word when the first word is presented in the upcoming test.

When you are ready to begin the restudy phase, click next.

[Word-Pairs Cycle Two Presented]

Now you have 2 minutes to solve as many of the following maths problems as possible. Please enter your answers in the correct order in the box below. Separate each answer with a comma (,).

[Filler Task Two Presented]

You will now undergo the final test to see how many of the word pairs you can remember. During the test, the first word will be displayed and you will be asked to type the second word of the pair. This part of the study is not timed, and you may take as long as you like. If you do not know the answer, you may type "X" to indicate this. Once you are satisfied with your answer, press NEXT to move to the next screen.

[Cued Recall Task Presented]

Thank you for your participation. Your time is very much appreciated.